# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-01-0188*



| 1. REPORT DATE *(DD-MM-YYYY)* | 2. REPORT TYPE | 3. DATES COVERED *(From - To)* |
|---|---|---|
| 2002 | Journal Article | |

**4. TITLE AND SUBTITLE**
WEIGHING THE EVIDENCE OF ECOLOGICAL RISK FROM CHEMICAL CONTAMINATION IN THE ESTUARINE ENVIRONMENT ADJACENT TO THE PORTSMOUTH NAVAL SHIPYARD, KITTERY, MAINE, USA

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHORS**
Robert K. Johnston, Wayne R. Munns, Jr., Patti Lynne Tyler, Patty Marajh-Whittemore, Kenneth Finkelstein, Kenneth Munney, Fred T. Short, Ann Melville, and Simeon P. Hahn

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
SSC San Diego
53560 Hull Street
San Diego, CA 92152-5001

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

**10. SPONSOR/MONITOR'S ACRONYM(S)**

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION/AVAILABILITY STATEMENT**
Approved for public release; distribution is unlimited.

**13. SUPPLEMENTARY NOTES**


**14. ABSTRACT**
In characterizing ecological risks, considerable consensus building and professional judgments are required to develop conclusions about risk. This is because how to evaluate all the factors that determine ecological risk is not well defined and is subject to interpretation. Here we report on the application of a procedure to weigh the evidence of ecological risk and develop ecological risk of chemical contamination in nearshore areas adjacent to the Portsmouth Naval Shipyard, located at the mouth of the Great Bay Estuary, New Hampshire and Maine, USA. Measures of exposure and effect were used to interpret the magnitude of risk to the assessment endpoints of pelagic species, epibenthic species, the benthic community, eelgrass plants, the salt marsh community, and avian receptors. The evidence of chemical exposure from water, sediment, and tissue and the evidence of biological effects to representative pelagic, epibenthic, benthic, eelgrass, salt marsh, and avian species were weighted to characterize ecological risk. Individual measures were weighted by the quality and reliability of their data and risk was estimated from the preponderance, magnitude, extent, and strength of causal relationships between the data on exposure and effects. Relating evidence of risk to hypothesized pathways of exposure made it possible to estimate the magnitude of risk from sediment and water and express the confidence associated with the findings. Systematically weighting the evidence of risk rendered conclusions about risk in a manner that was clearly defined, objective, consistent, and did not rely solely on professional judgment.

Published in *Environmental Toxicology and Chemistry*. Vol. 21, No. 1, pp. 182–194, 2002.

**15. SUBJECT TERMS**
Estuarine
Ecological risk assessment
Exposure assessment
Effects assessment
Risk characterization

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | Robert K. Johnston, Code 2375 |
| U | U | U | UU | 13 | 19B. TELEPHONE NUMBER *(Include area code)* (619) 553-2213 |

Environmental Toxicology and Chemistry, Vol. 21, No. 1, pp. 182–194, 2002
Printed in the USA
0730-7268/02 $9.00 + .00

*Hazard/Risk Assessment*

# WEIGHING THE EVIDENCE OF ECOLOGICAL RISK FROM CHEMICAL CONTAMINATION IN THE ESTUARINE ENVIRONMENT ADJACENT TO THE PORTSMOUTH NAVAL SHIPYARD, KITTERY, MAINE, USA

ROBERT K. JOHNSTON,*† WAYNE R. MUNNS, JR.,‡ PATTI LYNNE TYLER,§ PATTY MARAJH-WHITTEMORE,‖ KENNETH FINKELSTEIN,# KENNETH MUNNEY,†† FRED T. SHORT,‡‡ ANN MELVILLE,§§ and SIMEON P. HAHN‖‖
†Marine Environmental Support Office, Space and Naval Warfare Systems Center D3621, San Diego, California 92152-6326, USA
‡U.S. Environmental Protection Agency, National Health and Environmental Effects Research Laboratory, Narragansett, Rhode Island 02882-1154
§Office of Ecosystem Assessment, U.S. Environmental Protection Agency, Lexington, Massachusetts 02421
‖Waste Management Division, U.S. Environmental Protection Agency, Region I, Boston, Massachusetts
#National Oceanic and Atmospheric Administration, Boston, Massachusetts 02114-2023, USA
††U.S. Fish and Wildlife Service, 22 Bridge Street, Concord, New Hampshire 03301
‡‡Jackson Estuarine Laboratory, University of New Hampshire, Durham, New Hampshire 03824-3427, USA
§§Bureau of Oil and Hazardous Materials, Maine Department of Environmental Protection, Augusta, Maine 04333-0017, USA
‖‖Biological Sciences, Naval Facilities Engineering Command, Lester, Pennsylvania 19113-2090, USA

(*Received* 16 *January* 2001; *Accepted* 30 *May* 2001)

**Abstract**—In characterizing ecological risks, considerable consensus building and professional judgments are required to develop conclusions about risk. This is because how to evaluate all the factors that determine ecological risk is not well defined and is subject to interpretation. Here we report on the application of a procedure to weigh the evidence of ecological risk and develop conclusions about risk that will incorporate the strengths and weaknesses of the assessment. The procedure was applied to characterize ecological risk of chemical contamination in nearshore areas adjacent to the Portsmouth Naval Shipyard, located at the mouth of the Great Bay Estuary, New Hampshire and Maine, USA. Measures of exposure and effect were used to interpret the magnitude of risk to the assessment endpoints of pelagic species, epibenthic species, the benthic community, eelgrass plants, the salt marsh community, and avian receptors. The evidence of chemical exposure from water, sediment, and tissue and the evidence of biological effects to representative pelagic, epibenthic, benthic, eelgrass, salt marsh, and avian species were weighed to characterize ecological risk. Individual measures were weighted by the quality and reliability of their data and risk was estimated from the preponderance, magnitude, extent, and strength of causal relationships between the data on exposure and effects. Relating evidence of risk to hypothesized pathways of exposure made it possible to estimate the magnitude of risk from sediment and water and express the confidence associated with the findings. Systematically weighing the evidence of risk rendered conclusions about risk in a manner that was clearly defined, objective, consistent, and did not rely solely on professional judgment.

**Keywords**—Estuarine Ecological risk assessment Exposure assessment Effects assessment Risk characterization

## INTRODUCTION

### *Overview*

Ecological risks are characterized by using data from field and laboratory studies. Results from multiple measures of environmental condition must be synthesized and reconciled. Weighing multiple lines of evidence to develop conclusions has been used in many ecological risk [1,2] and sediment quality studies [3,4]. The weight of evidence provides a means of developing conclusions that are based on all the available data. Generally, equal weight is given to each line of evidence. However, when lines of evidence are ambiguous or in conflict, final estimates of risk and harm require considerable professional judgment. Previously, the Massachusetts weight-of-evidence workgroup developed a methodology for assigning different weights to the measurement endpoints and recommended a weight-of-evidence procedure for characterizing ecological risk [5]. In this approach, each measurement endpoint is weighted based on attributes of data quality, strength of association to the assessment endpoint, and study design. Then the magnitude of response obtained for each measurement endpoint is summarized and conclusions about risk to the assessment endpoints are formulated based on the concurrence among the weighted measurement endpoints [5].

Here we present a case study on the use of the workgroup's approach for assessing ecological risks from the release of hazardous chemicals [6,7] from the Portsmouth Naval Shipyard [8,9] located on Seavey Island, Maine, USA (Fig. 1). The approach was used to examine the strengths and weaknesses of the various measurements and assign an endpoint weight to each measure by evaluating the strength of association between the assessment endpoint and the measurement, the quality of its data, and the design of the study. The conclusions about risk were based on the amount of evidence (preponderance), the degree of evidence for an exposure or an effect (magnitude), the spatial extent of the measured effects, and the link between exposure and effects (causation). In the process, we improved and refined the procedures advocated by the [illegible] and reached conclusions about risk that could

* To whom correspondence may be addressed (johnston@spawar.navy.mil).

Contribution NHEERL-NAR-2148, U.S. Environmental Protection Agency, contribution 345, Jackson Estuarine Laboratory, Durham, New Hampshire.

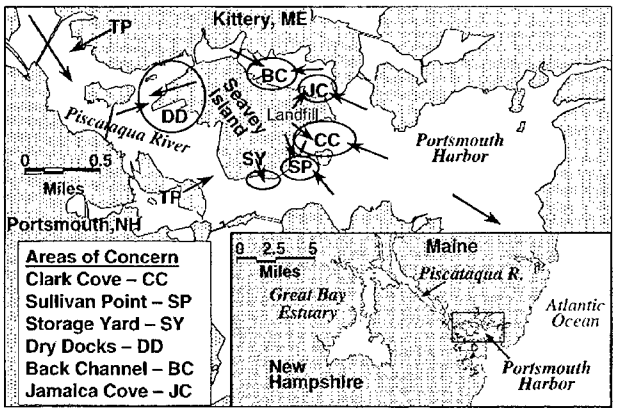


Fig. 1. Conceptual model for the lower Piscataqua River, New Hampshire, USA, showing the location of the Portsmouth Naval Shipyard (on Seavey Island, ME), sewage treatment plants (TP), the areas of concern (circles), and hypothesized waterborne transport in the estuarine system (arrows).

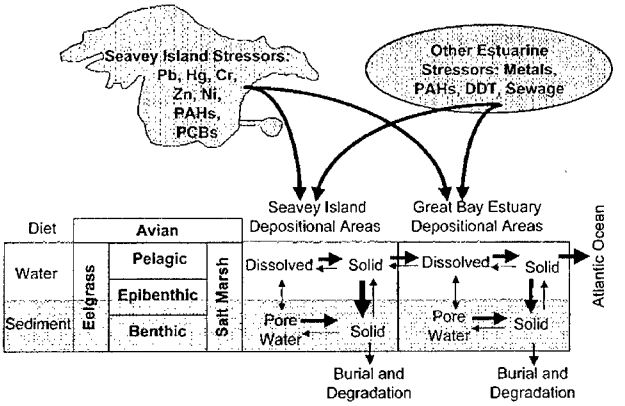


Fig. 2. Details of conceptual model showing release of stressors, accumulation in depositional areas, settling, geochemical partitioning between dissolved and solid phases, burial and degradation, loss to the ocean, and the relationship of the assessment endpoints to exposure from sediment, water, and diet.

be used by the regulatory community and the interested public and that helped the Navy develop cleanup strategies. The procedures, supporting data, and technical information are included in the ecological risk assessment submitted by the Navy [7].

*Background*

The estuarine system formed by the Great Bay and Piscataqua River extends 20 to 25 miles into New Hampshire, USA (Fig. 1). The estuary is fed by seven rivers, and more than 220,000 people live within the watershed [10]. The estuary is flushed extremely well [5,11,12]. Flushing times for the lower estuary have been estimated at 1.6 to 2.2 d [12], and for the headwaters, about 18 d [12]. The strong tidal currents scour the bottom of the main channel and leave a substrate of gravelly sand [13], but the weaker currents in coves and channels deposit sediment [13]. The bottom of the estuary is covered with glacial tills, stratified deposits, and glacial marine sediments. These sediments accumulate wherever river flow is reduced. Deposits of muds and muddy sands are present in coves and confined channels of the lower estuary [13], including Clark and Jamaica Coves, the Back Channel, and areas of concern around Seavey Island, Maine, USA (Fig. 1). These depositional areas provide habitat for a wide variety of fish and invertebrates, including winter flounder, lobster, blue mussels, eelgrass, and waterfowl [14]. Small marshes with well-developed substrata of peat are also found within some of the depositional areas [15].

Seavey Island has been used as a navy yard since before the Revolutionary War. The Navy's first submarines were built at the Portsmouth Navy Yard, where more than 20,000 men and women worked during the height of World War II [16]. Past practices at the shipyard resulted in the release of wastes containing metals, cyanide, polychlorinated biphenyls (PCBs), phenols, oils, and grease into the estuary [16]. From 1945 to 1978, hazardous wastes were disposed in a landfill created by filling tidal flats with materials, which included sludge, solvents, asbestos, blasting grit, incinerator ash, waste oils, and spoils dredged from near the dry docks [17]. A storage yard on the south shore of the island was also contaminated with Pb, Cu, Zn, PCBs, and other semivolatile compounds [18]. Through ongoing cleanup activities at the shipyard, further contamination of the estuary from the shipyard's solid-waste management units will be prevented [19].

*Conceptual model*

The approach recommended by the U.S. Environmental Protection Agency (U.S. EPA) Risk Assessment Forum [20,21] requires two types of information to characterize ecological risks from contamination, i.e., chemical exposure in environmental media (river water, sediments, and biota) and relations between exposures (doses) and measurable ecological effects. We characterized the ecological risk by relating measures of chemical contamination to assessment endpoints in the estuary, where assessment endpoints are defined as the environmental conditions or processes that we desire to protect [22]. The assessment endpoints consisted of the health and vitality of pelagic species, epibenthic species, the benthic community, eelgrass plants, the salt marsh community, waterfowl, and birds of prey. In order to relate levels of exposure to potential effects on the assessment endpoints, receptors of concern (species or communities of species that can be evaluated at the site) in the Great Bay Estuary were identified for each assessment endpoint [7].

In order to assess the ecological effects of contaminants released from the shipyard, a conceptual model was developed [6,7] to predict their behavior after being released. The first tier of the conceptual model describes the waterborne transport of chemicals released into the estuary (Fig. 1). Important sources of chemical pollution of Portsmouth Harbor included the shipyard, the sewage treatment plants, up-estuary sources of Cr, Ni, and PAHs, and runoff from nonpoint sources. Contaminants that enter the river will be mixed quickly into the water column. Chemicals that dissolve should then be diluted and flushed from the system, but those that persist and attach themselves to particles will accumulate in the areas where sediments are deposited. These depositional areas will accumulate contaminants from all sources in the estuary (Fig. 2). Once chemicals become associated with the sediment, they may bind to the solid phase, partition to the pore water, or be resuspended by tidal currents. Bioturbation, biotransformation, and bioaccumulation may then redistribute them. Chemicals will be buried wherever sediments accumulate fast enough. Aquatic organisms can be exposed to chemicals present in the water column, sediment, pore water, and prey (Fig. 2).

Table 1. The scheme used to interpret the results of measures of exposure and effects

| Type of measure | Degree of response | Interpretation | Value ($M_i$) |
|---|---|---|---|
| Exposure | ≤Reference condition or below conservative benchmark concentration | Negligible exposure | 0 |
| | >Reference condition | Low exposure | 1 |
| | Statistically > reference concentration | Elevated exposure | 2 |
| | >Conservative benchmark concentration | High exposure | 3 |
| | >Nonconservative benchmark concentration | Adverse exposure | 4 |
| Effect | Similar to reference or control condition or below ecologically relevant threshold | No effect | 0 |
| | Worse than reference or control condition but not statistically different | Potential effect | 1 |
| | Statistically worse than reference or control condition | Probable effect | 2 |

The fact that persistent chemicals are trapped close to the organisms that live in depositional habitats means these areas pose a greater ecological risk than nondepositional areas do. We focused on ecological risks in nearshore depositional areas around Seavey Island (areas of concern) because they would be most likely to accumulate contaminants from the shipyard (Fig. 1). We offer Clark Cove (Fig. 1) as an example of how we weighed the evidence of risk to each area of concern. Clark Cove was the major focus of the ecological risk assessment; more data on exposure and effects were obtained here than in any of the other areas studied. Because the procedure for the other areas of concern was similar, we save space by omitting those analyses, whose details are contained in the Navy's ecological risk assessment [7].

*Chemicals of concern*

In order for contaminants in the estuary to be linked with the disposal sites on the shipyard, there must be a plausible route from the waste sites to the estuary. Even though information on past releases from the shipyard is incomplete, it is certain that the shipyard has contributed pollutants to the estuary. Because the contaminants in the estuary could have come from other sources as well, it was necessary to determine which chemicals in the estuary were elevated and which of those could have come from the shipyard. Chemicals that exceeded background soil concentrations for Seavey Island at the disposal sites, had a migratory pathway to the estuary, and showed evidence of a spatial gradient from the shipyard or exceeded thresholds of toxicity in sediment, water, and tissue samples from the estuary were identified as contaminants of concern for the risk assessment [7]. They were Pb, Hg, Cu, Cr, Ni, Zn, Ag, As, Cd, polycyclic aromatic hydrocarbons (PAHs; individually and summed together), PCBs (individual congeners and total PCB), and the pesticide compounds (individually and summed together as tDDx) dichlorodiphenyl trichloroethane, dichlorodiphenyl dichloroethane, and dichlorodiphenyl dichloroethylene [7].

## METHODS

The weight-of-evidence analysis consisted of the following steps: Endpoint weights were objectively assigned to each measure of exposure and effect. The endpoint weight was based on the strength of the relationship to the assessment endpoint, data quality, and study design (Appendix).

The outcomes of the measures were interpreted based on whether the result added weight to the conclusion of risk or no risk (Table 1). Summary tables for each assessment endpoint and area of concern were constructed that contained all the information available to evaluate risk.

Definitions of risk were developed to interpret the results of the exposure and effects information (Table 2).

Scatter plots of the outcomes of the exposure and effect measures were plotted versus their corresponding endpoint weights (Fig. 3). This allowed the results obtained for each assessment endpoint to be visualized.

A centroid was calculated that consisted of a weighted average of the outcomes (weighted by their endpoint weights).

The interpretation of risk and confidence in conclusions were summarized for each assessment endpoint.

The evidence of risk was related to hypothesized pathways of exposure to estimate the magnitude of risk from sediment and water and express the confidence associated with the findings. The details of these procedures are provided below.

Endpoint weights were assigned to each measure of exposure and effect to reflect the reliability and usefulness of the measure to assess risk to the assessment endpoint. For each of the exposure and effects measures [6,7], the data quality, the strength of its association to the assessment endpoint, and the study design were evaluated (Appendix). The weighting procedure consisted of scoring the attributes of each measure as low, medium, or high, depending on how well the measurement data related to assessing stressor levels or ecological damage. Based on the scores assigned to the three categories of attributes, the endpoint weight ($W_i$) for each measurement

Table 2. Interpretation of exposure and effect evidence in determining risk

| Evidence of effect | Evidence of exposure | | | | |
|---|---|---|---|---|---|
| | Negligible | Low | Elevated | High | Adverse |
| No | Negligible | Negligible | Low | Low | Intermediate |
| Potential | Negligible | Low | Intermediate | Intermediate | High |
| Probable | Low | Low | Intermediate | High | High |

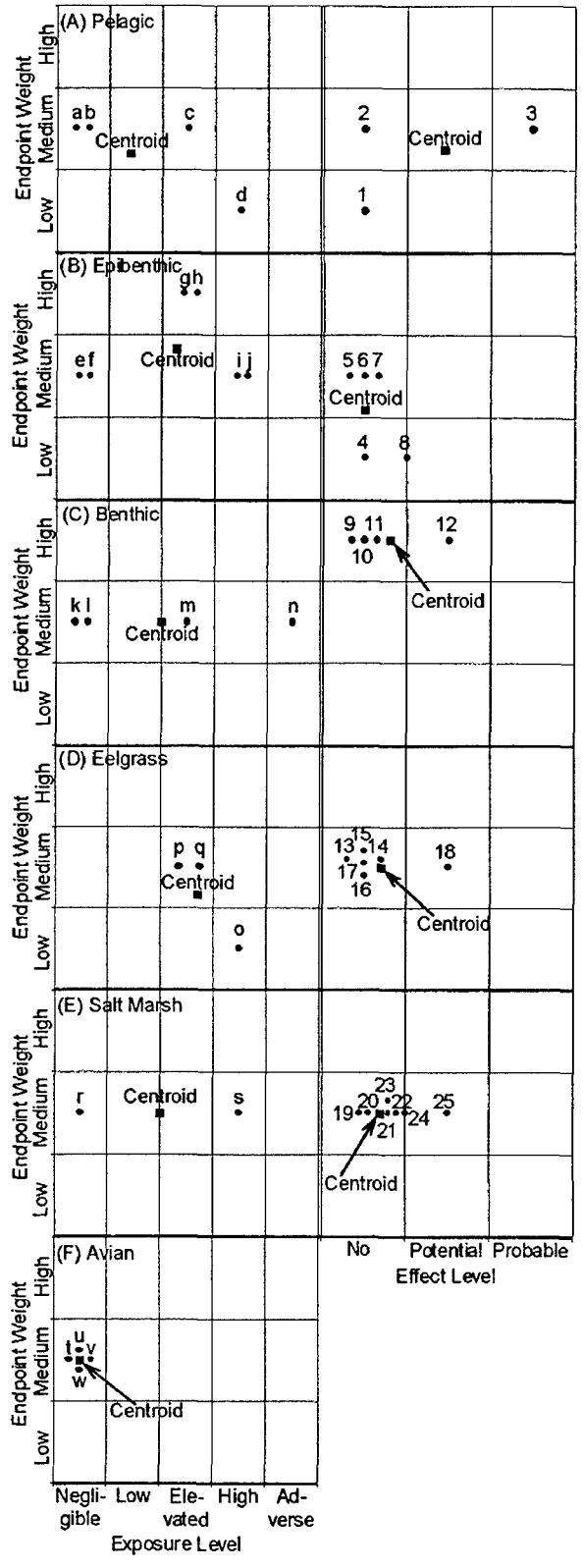


Fig. 3. The outcome for measures of exposure and effects to (**A**) pelagic species, (**B**) epibenthic species, (**C**) benthic community, (**D**) eelgrass plants, (**E**) salt marsh community, (**F**) avian receptors in Portsmouth Harbor, and the centroid suggested by the weighted average of the measures. See Table 5 for definitions of symbols used to represent the outcomes of measures of exposure and effects.

($i$) was determined by our professional judgment. The possible endpoint weights were low (1), medium (2), or high (3). The endpoint weight represented our confidence in using the measurement to infer harm to the assessment endpoint.

To interpret the outcomes of exposure and effects measurements, site-specific responses were compared to biologically based benchmarks or to measurements obtained at reference sites (or under controlled conditions) and to the expected range of variation based on professional judgment (Table 1). Benchmarks are concentrations in sediment, water, or tissue that may give rise to biological responses when exceeded. Typically, measures of exposure were evaluated by comparing with benchmarks or reference conditions, and measures of effect were evaluated by comparing them with reference (control) responses or with the expected range of variation. Chemical concentrations below conservative benchmarks, such as the no-observed-effect concentration (NOEC) or effects range low (ERL) [23] were interpreted as negligible exposure. Concentrations above the NOEC were interpreted as high exposure and concentrations above a nonconservative benchmark such as lowest-observed-effect concentration (LOEC) or effects range median (ERM) [23] were interpreted as adverse exposure. When benchmarks were not available for an exposure measure (e.g., chemical residues in eelgrass plants), the interpretation was negligible, low, or elevated based on comparing the result with reference areas (Table 1). For measures of effect, the interpretation was based on control response (e.g., sediment toxicity) or responses obtained from reference areas (e.g., salt marsh species richness).

The evidence found for exposure and effects defined the risk levels (Table 2). The more evidence of exposure and effects, the greater the risk, while evidence of exposure or effect without evidence of the other suggested lesser risk. Negligible risk means that the data suggested no impacts and that there was a general lack of evidence of exposure or effects. Low risk means that the data suggested limited impact but there was little correspondence between measures of exposure and effect. Intermediate risk means that the data suggested there were potential impacts and that measures of effect were associated with measures of exposure. High risk means that the data indicated large and persistent impacts and that there was a direct relationship between measures of exposure and effect.

To visualize the weight of evidence, scatter plots of the outcomes of exposure and effects ($M_i$) were plotted versus their corresponding endpoint weights ($W_i$). The scatter plots were used to evaluate the weight of evidence of risk for each assessment endpoint (Fig. 3). A centroid weighted by the endpoint weights was calculated to help visualize the preponderance of the results. The centroid was used to aid in interpreting the balance of exposure and effects information suggested by the data. Measures with higher weight would tend to draw the centroid in their direction. The centroid was plotted as ($X_w$, $Y$), where $Y$ was the arithmetic average of the endpoint weights ($Y = (\Sigma\ W_i)/n$) and $X_w$ the weighted average of the exposure or effect outcomes ($X_w = (\Sigma\ (M_i \cdot W_i))/\Sigma\ W_i$). For clarity, individual measures on the scatter plots were identified and the centroid's location was used to guide the interpretation of risk. If the centroid fell on a boundary between two outcomes, the most conservative interpretation was chosen.

Based on the evidence of exposure and effects, the magnitude of risk ($R_i$) for each assessment endpoint ($i$) was obtained from Table 2. The confidence level ($C_i$) in the risk estimate was based on the average of the endpoint weights

Table 3. Weights ($WM_i$) used for calculating magnitude of risk from medium and confidence in conclusions; the weight was attributed between the two media to reflect the assumed predominant route of exposure; because two routes of exposure were evaluated, the maximum weight possible was 2

| Assessment endpoint | Surface water $WM_{water}$ | Sediment $WM_{sediment}$ |
|---|---|---|
| Pelagic | 2 | 0 |
| Epibenthic | 1 | 1 |
| Benthic | 0 | 2 |
| Eelgrass | 1 | 1 |
| Salt marsh | 1 | 1 |

and the extent of agreement between the various estimates. For example, a tight scatter with high weight would increase the confidence, while a broad scatter with lower weight would decrease confidence. If necessary, we used professional judgment to qualify our conclusions.

*Attributing risk to environmental media*

The risks to assessment endpoints were attributed to the exposure media (estuarine water and sediment) to relate risk back to possible cleanup options for the site. The risk from the media ($R_M$) and the confidence in the conclusion ($C_M$) was calculated as the weighted average of the risks to the assessment endpoints. Because individual assessment endpoints may preferentially offer information on exposures from surface water or from sediment, the endpoints were weighted by the degree that we expected them to have been influenced by the exposure media (Fig. 2, Table 3). For example, because the measures used for the benthic assessment endpoint provided more information about exposure from sediment than the measures used for the pelagic assessment endpoint, the benthic assessment endpoint was weighted higher than the pelagic assessment endpoint when assessing risk from exposure to sediment (Table 3). Conversely, the measures used for the pelagic assessment endpoint provided the most information about exposures from surface water, and the measures used for the eelgrass, epibenthic, and salt marsh assessment endpoints provided information on exposures from both sediments and surface waters.

To convert the description of risk and confidence to a numerical value, values were assigned to the levels of risk and confidence (Table 4). The magnitude of risk ($R_i$) and the confidence level ($C_i$) for each assessment endpoint were assigned a numeric value (Table 4) and weighted (Table 3) to estimate the risk caused by exposure to sediment and water. The magnitude of risk from a medium then became $R_M = (\Sigma\ R_i \cdot WM_i)/\Sigma\ WM_i$ and the confidence $C_M = (\Sigma\ C_i \cdot WM_i)/\Sigma\ WM_i$, where $R_i$ is the magnitude of risk for an assessment endpoint $i$, $C_i$ is the confidence in the risk for the same endpoint, $WM_i$ is the weight used to evaluate the risk (Table 3), and cutoff values (Table 4) were used to determine the magnitude of risk from exposure medium ($R_M$) and confidence in conclusion ($C_M$). If any of the assessment endpoints did not apply to a specific area of concern, we excluded them from the calculation. By relating the risk to the exposure media, we were able to assess the degree to which sediment and water contributed to the various risks.

Table 4. Numeric values assigned to the magnitude of risk to assessment endpoint ($R_i$) and confidence in conclusion ($C_i$) and cut-off values for determining magnitude of risk from exposure medium ($R_M$) and confidence in conclusion ($C_M$)

| Magnitude of risk to assessment endpoint | Numeric value[a] ($R_i$) | Cut-off value[b] | Magnitude of risk from exposure medium ($R_M$) |
|---|---|---|---|
| Negligible | 0 | <0.50 | Negligible |
| Low | 1 | <1.25 | Low |
| Intermediate | 2 | <2.00 | Intermediate |
| High | 3 | ≤3.00 | High |
| **Confidence in conclusion ($C_i$)** | **Numeric value[a]** | **Cut-off value[b]** | **Confidence in conclusion ($C_M$)** |
| Low | 1 | <1.667 | Low |
| Medium | 2 | <2.333 | Medium |
| High | 3 | ≤3.000 | High |

[a] Numeric value is used to convert qualitative statement to a quantitative value (e.g,. negligible to 0) for use in calculating the weighted average for $R_M$ and $C_M$.

[b] Cut-off value is used to convert the quantitative value derived for $R_M$ and $C_M$ into a qualitative statement (e.g., $R_M$ = 1.6 to intermediate).

## RESULTS

*Endpoint weights*

An overall endpoint weight for each measure was determined based on the qualitative scores (low, medium, and high) assigned to data quality, strength of association, and study design (Table 5). We judged data quality to be particularly important, and so low-quality data were not used in the risk assessment. Data from the measurements were reported in [6] and [7].

For assessing effects on pelagic receptors, data on phytoplankton biomass (estimated from concentrations of chlorophyll *a* and phaeopigments), toxicity to fertilization of sea urchin (*Arbacia punctulata*), scope for growth of deployed mussels (*Mytilus edulis*), and size, abundance, and spleen histopathology of winter flounder (*Plueronectes americanus*) were used. The overall weights for biomass of phytoplankton and abundance and size of flounder were low (Table 5) because these measures may be affected by many other factors besides chemical stressors (which makes the basis for inferring harm to pelagic species weak) and because the sampling design did not adequately account for temporal and spatial variability. For toxicity to sea urchins, data quality was weighted medium because the 48-h holding time for the toxicity samples was exceeded; study design was also weighted medium because it was tested during only one sampling period; and strength of association was high because toxicity to sea urchin larvae implies a potential impact to pelagic species that broadcast their sperm and larvae to the water column. This endpoint was given an overall weight of medium. Scope for growth in deployed mussels was weighted medium overall because it may be insensitive to chemicals while correlating strongly with somatic growth. Unfortunately, scope for growth was measured only once, and the reference station used to evaluate the results may not have represented the areas of concern. For measures of spleen histopathology, there is a good correlation of contaminants to pathological effects, but abnormal spleen pathology was observed in winter flounder collected from both

Portsmouth Harbor and the Gulf of Maine reference area [7]. The measure of spleen histopathology was assigned an endpoint weight of medium (Table 5).

For exposure of pelagic receptors, the concentration of chemicals measured in estuarine surface water was weighted medium overall, with its data quality weighted high and its strength of association and study design medium. The concentration of chemicals measured in seep samples was rated low overall, with high data quality, but the strength of association and study design were low because seeps are diluted rapidly as they enter the harbor and the seeps were not sampled frequently enough. Accumulation of contaminants in tissues of deployed mussels was weighted medium overall, with a high strength of association and high and medium for data quality in the two cases. The study design was weighted low because the two deployments of the caged mussels used different stations and different durations (one month and three months). Residues of contaminants in flounder liver and fillet tissues were weighted medium because the quality of the data was high, there was a medium strength of association between the accumulation of contaminants in flounder tissues and exposure to pelagic receptors, but the study design (low) lacked the ability to distinguish between spatial, temporal, and natural variations [7].

Measurements of winter flounder's abundance, size, histopathology, and tissue residues and adult lobster abundance, size, and tissue residues were evaluated for Portsmouth Harbor as a whole. While demersal fish and lobsters can indicate the levels of environmental pollution well because they live relatively long, they associate closely with sediment, and they feed mainly on benthic invertebrates, they may not stay in close proximity to the site, resulting in uncertainty in relating results back to contamination originating from the shipyard.

For assessing effects on epibenthic receptors, density of the lobster *Homarus americanus,* density, shell length, and condition index of the indigenous mussel *Mytilus edulis,* and biomass of the fucoid *Ascophyllum nodosum* were evaluated. The strength of association for lobster and fucoid density was rated low (Table 5) because there is not adequate data to link decrease in density to elevated chemical concentrations in environmental media [7]. The strength of association for mussel density, condition index, and length was rated medium because mussels are sessile and there is a plausible link between environmental contamination and effects to these measures [7]. The study design was weighted low because the sampling intervals were not sufficient to determine whether differences were statistically significant. Natural stochasticity and the experimental/analytical variability in these measures make the situation worse.

For assessing exposures of epibenthic receptors, chemical concentrations in estuarine surface water, seep water, and tissues of lobsters, fucoids, and mussels were evaluated. The data on chemical concentrations in juvenile lobsters were assigned a high weight because tagging studies showed that juveniles remained in the proximity of specific depositional areas (while adults migrated over long distances), the sampling for juvenile lobsters was targeted on areas of known contamination near the shipyard and uncontaminated reference areas, and the sample size was sufficient to detect statistically significant differences [7]. The lack of benchmark concentrations for tissue residues in lobster, tissues in fucoids, and sediments increased the uncertainty for measures of epibenthic exposure.

For assessing effects on the benthic community, the density, richness, and evenness of benthic infauna and toxicity to amphipods (*Ampelisca abdita*) were evaluated. For assessing exposure, concentrations in bulk sediments, enrichments of metals in sediments relative to crustal ratios [24,25], differences between concentrations of acid-volatile sulfides (AVS) and simultaneously extracted metals (SEM) [26], and pore water concentrations of organic compounds predicted by equilibrium partitioning [27] were evaluated. Because these measures were less uncertain than those for the other assessment endpoints, we weighted them higher (Table 5). Even so, benthic organisms may be affected by stresses other than chemicals (e.g., enriched nutrients, type of sediment substrate, intraspecies competition, and patchiness). Sources of uncertainty included nonequilibrium partitioning for concentrations in pore water and lack of seasonal data for AVS. Although benchmark concentrations of chemicals in sediments are generally available, causal relationships between elevated concentrations and composition of benthic communities remain unclear [3–5].

For assessing effects on eelgrass (*Zostera marina*), its morphology (length and biomass of leaves and roots/rhizomes), its density of reproductive and vegetative shoots, its number of leaves per shoot, and the spatial distribution of its beds were evaluated. To assess its exposure to chemicals, chemical concentrations in its leaves, tissues of roots/rhizomes, and bulk sediment from its beds were evaluated. Most of the measures of effects on eelgrass were weighted medium (and one of six low) (Table 5) because benchmark concentrations for eelgrass are not available, because effects have not been correlated with contamination, and because the scientific basis for inferring environmental harm from measurements of eelgrass is still weak. Although chemicals in eelgrass are biologically mediated, data are lacking that relate concentrations to eelgrass effects.

For assessing effects to salt marshes, measures of salt marsh cord grass (*Spartina* spp.) cover, cover of other vascular plants, morphology of *Spartina* (height and density of stems, percentage of reproductive stems, and biomass above ground), number of animal taxa, abundances of amphipods and mollusks, and ratio of live:dead shells of snails (*Littorina littorea*) were evaluated. For each marsh, these measures were evaluated in areas of low marsh dominated by tall *Spartina alterniflora,* middle marsh dominated by short *S. alterniflora,* and high marsh dominated by *Spartina patens.* Chemical concentrations in *Spartina* leaves and bulk sediment in each marsh were evaluated for measures of exposure. Although most of the salt marsh measures incorporated community-level effects, the salt marsh study was descriptive [15] and effects observed had to be considered as potential only, with further study being needed to check their significance. The weights assigned to the exposure endpoints also had to be reduced because of the lack of benchmark effects on salt marsh plants and the fact that only *Spartina* leaves, not roots, were analyzed chemically (Table 5).

For assessing exposure to the avian receptors, modeled dietary exposure to black ducks (omnivores), Canada geese (herbivores), herring gulls (carnivores), and ospreys (piscivores) were evaluated. Because these birds can utilize the entire lower estuary, exposures for Portsmouth Harbor were evaluated as a whole by using the maximum exposures from food (prey and plants) and sediment in our calculations. Contaminant uptake by birds from estuarine waters was limited to dermal contact and was assumed to be negligible. The dietary-exposure model also assumed that the reference values for toxicity

Table 5. The assessment endpoints, the measure of exposure or effects, and the endpoint weight score assigned for data quality, strength of association with the assessment endpoint, study design, and endpoint weight. Measures were assigned an endpoint weight score of high (H), medium (M), or low (L) relative to their ability to assess harm to the assessment endpoint. Professional judgment was used to determine the endpoint weight based on the scores for data quality, strength of association, and study design. The symbols used in Figure 3 are also noted

| Assessment endpoint measure | Endpoint weight score: Data quality | Strength of association | Study design | Endpoint weight ($W_i$) | Figure 3 symbol |
|---|---|---|---|---|---|
| Pelagic community: measures of effect | | | | | |
| Phytoplankton biomass | H | L | L | L (1) | 1 |
| Scope for growth in mussels (*Mytilus edulis*) deployed for 28 d, fall 1991 | H | M | L | M (2) | 2 |
| Sea urchin (*Arbacia punctulata*) fertlization after sperm cells were exposed for 1 h to bulk water collected from the site | M | H | M | M (2) | 3 |
| Winter flounder (*Pleuronectes americanus*) abundance and size for Portsmouth Harbor as a whole | H | L | L | L (1) | |
| Winter Flounder (*P. americanus*) spleen histopathology for Portsmouth Harbor as a whole | H | M | L | M (2) | |
| Pelagic community: measures of exposure | | | | | |
| Estuarine surface water concn. | H | M | M | M (2) | a |
| Deployed *M. edulis* tissue concn. after 28 d deployment, fall 1991 | H | H | L | M (2) | b |
| Deployed *M. edulis* tissue concn. after 90 d deployment, fall 1993 | M | H | L | M (2) | c |
| Seep water contaminant concn. | H | L | L | L (1) | d |
| Winter flounder (*P. americanus*) liver tissue concn. for Portsmouth Harbor as a whole | H | M | L | M (2) | |
| Winter flounder (*P. americanus*) fillet tissue concn. for Portsmouth Harbor as a whole | H | M | L | M (2) | |
| Epibenthic community: measures of effect | | | | | |
| Lobster (*Homarus americanus*) density | H | L | L | L (1) | 4 |
| Indigenous *M. edulus* density | H | M | L | M (2) | 5 |
| Indigenous *M. edulis* shell length | H | M | L | M (2) | 6 |
| Indigenous *M. edulis* condition index | H | M | L | M (2) | 7 |
| Fucoid algae (*Ascophyllum nodosum*) biomass | H | L | L | L (1) | 8 |
| Epibenthic community: measures of exposure | | | | | |
| Estuarine surface water concn. | H | M | M | M (2) | e |
| *A. nodosum* tissue concn. | H | M | L | M (2) | f |
| Juvenile *H. americanus* tail and claw tissue concn. | H | H | M | H (3) | g |
| Juvenile *H. americanus* hepatopancreas tissue concn. | H | H | M | H (3) | h |
| Adult *H. americanus* tail and claw tissue concn. | H | M | L | M (2) | |
| Adult *H. americanus* hepatopancreas tissue concn. | H | M | L | M (2) | |
| Seep water concn. | H | M | L | M (2) | i |
| Indigenous *M. edulis* tissue concn. | H | M | M | M (2) | j |
| Benthic community: measures of effect | | | | | |
| Amphipod (*Ampelisca abdita*) mortality after 10 d exposure to bulk sediment collected from the site | H | H | M | H (3) | 9 |
| Benthic community richness | H | H | M | H (3) | 10 |
| Benthic community density | H | H | M | H (3) | 11 |
| Benthic community evenness | H | H | M | H (3) | 12 |
| Benthic community: measures of exposure | | | | | |
| Concentration of acid volatile sulfide (AVS) μmol/g dry wt minus concn. of simultaneously extracted metal (SEM) μmol/g dry wt (AVS-SEM) [26] | H | M | L | M (2) | k |
| Pore water toxicity predicted using equilibrium partitioning assumptions and compared with chronic water quality criteria or LC50 data [27] | H | M | M | M (2) | l |
| Metal enrichment estimated from the concn. of Al in the sample [24,25] | H | L | M | M (2) | m |
| Bulk sediment contaminant concn. | H | M | M | M (2) | n |
| Eelgrass (*Zostera marina*) plants: measures of effect | | | | | |
| *Z. marina* leaf morphology | H | M | M | M (2) | 13 |
| *Z. marina* root and rhizome morphology | H | M | M | M (2) | 14 |
| *Z. marina* vegetative shoot density | H | M | M | M (2) | 15 |
| *Z. marina* reproductive shoot density | H | M | L | M (2) | 16 |
| *Z. marina* ratio of leaves to shoots | H | L | M | L (1) | 17 |
| *Z. marina* spatial distribution | H | M | L | M (2) | 18 |
| Eelgrass (*Z. marina*) plants: measures of exposure | | | | | |
| Bulk sediment contaminant concn. | H | L | L | L (1) | o |
| *Z. marina* leaf tissue concn. | H | M | M | M (2) | p |
| *Z. marina* root and rhizome concn. | H | M | M | M (2) | q |

Table 5. Continued

| Assessment endpoint measure | Endpoint weight score: Data quality | Strength of association | Study design | Endpoint weight ($W_i$) | Figure 3 symbol |
|---|---|---|---|---|---|
| Salt marsh community: measures of effect | | | | | |
| *Spartina* spp. cover | H | M | M | M (2) | 19 |
| *Spartina* spp. morphology | H | M | M | M (2) | 20 |
| Amphipod abundance | H | M | M | M (2) | 21 |
| Mollusk abundance | H | M | M | M (2) | 22 |
| No. of animal taxa | H | M | M | M (2) | 23 |
| Cover of vascular plants other than *Spartina* spp. | H | M | M | M (2) | 24 |
| Ratio of live to dead gastropod (*Littorina littorea*) shells | H | L | M | M (2) | 25 |
| Salt marsh community: measures of exposure | | | | | |
| *Spartina* spp. leaf tissue concn. | H | M | M | M (2) | r |
| Bulk sediment contaminant concn. | H | L | M | M (2) | s |
| Avain receptors: measures of exposure | | | | | |
| Dietary exposure to herbivore–Canada goose (*Branta canadensis*) | H | M | M | M (2) | t |
| Dietary exposure to omnivore–black duck (*Anas rubripes*) | H | M | M | M (2) | u |
| Dietary exposure to piscivore–osprey (*Pandion haliaetus*) | H | M | M | M (2) | v |
| Dietary exposure to carnivore–herring gull (*Larus argentatus*) | H | M | M | M (2) | w |

(based on receptor-specific no-observed-adverse-effect levels—which were adjusted by body wt and uncertainty factors) applied to the receptors of concern [28], that all chemicals were assimilated 90%, that the selected food items comprised 100% of the diets, and that the receptors fed only in Portsmouth Harbor [7]. Incidental sediment ingestion was assumed to be 10%, which is conservative based on literature values of the same or similar species, and food ingestion rates were calculated based on species-specific formulas from the literature [28]. Exposure duration was assumed to be 12 months for the Canada goose, black duck, and herring gull that potentially overwinter at the site. The osprey was thought to leave the site in winter, and therefore exposure was based on a half-year exposure cycle. This approach was unlikely to underestimate exposures for avian consumers because most of them migrate, because no-observed-adverse-effect levels are usually far below the lowest-observed-effects levels for most contaminants, because assimilation efficiency is probably less than 90% for most chemicals, and because maximum concentrations in prey and incidental exposure to sediment were used in the model.

*Evidence of risk*

The weight of evidence for risk was evaluated by plotting the outcomes of exposure and effects measures (Figs. 3A through F). The outcomes to pelagic species from the measures of exposure showed evidence of high concentrations in seep samples, elevated exposure in mussel tissues after the fall 1993 deployment (90 d), negligible exposure in mussel tissues after the fall 1991 deployment (28 d), and negligible concentrations in water samples from the cove. The weight of evidence for pelagic receptors in Clark Cove provided for low exposure with medium weight (Fig. 3A). For measures of effect, toxicity to sea urchin fertilization indicated a probable effect, but phytoplankton biomass and the scope for growth of mussels deployed in the cove indicated no effect. Since the endpoint weight for phytoplankton biomass was low and the weights for mussel growth and sea urchin toxicity were medium, we concluded medium weight of potential effects to pelagic receptors (Fig. 3A).

Although water from the seeps would be quickly diluted as it entered the cove, we noted that the seeps were not well characterized. The agent of toxicity in the sea urchin test was unknown. This test might have been affected because its water samples had been held for too long (samples were collected September 13–17, 1991, and the tests were conducted October 8–9, 1991), which could either increase or decrease the observed toxicity [6]. Furthermore, the two deployments of caged mussels involved different stations, different sampling times, and different lengths of deployment. These differences may have created the differences in exposure suggested by the different outcomes (Fig. 3A).

The weight of evidence for exposure and effects to epibenthic species in Clark Cove indicated medium weight of elevated exposure but with no effect (Fig. 3B). One of two fucoid algae monitoring stations in Clark Cove had less biomass than the reference area, suggesting a potential effect (plotted as 8 in Fig. 3B). However, measures of mussel density, length, and condition index and lobster density were similar or greater than reference areas, suggesting no effect (Fig. 3B). The outcomes also showed high exposure from the chemicals in seep water and in tissues of indigenous mussels, elevated exposure from concentrations in tissues of juvenile lobster, and negligible exposure from the low concentrations of chemicals in cove water and tissues of fucoid algae (Fig. 3B).

Effects to the benthic community were assessed using highly weighted measures for density, richness and evenness of the benthic community, and sediment toxicity to amphipods. Although evenness of species may have been affected, the other measures seemed unaffected, and we concluded high weight of no effect (Fig. 3C). The weight of evidence for the measures of exposure to the benthic community was interpreted to mean medium weight of elevated exposure (Fig. 3C). This conclusion was based on measures of exposure that showed high concentrations of chemicals in bulk sediments, metals in sediments enriched relative to the crust, negligible exposure from AVS–SEM, and predicted toxicity in pore waters (Fig. 3C). Uncertainties arose from the lack of data on seasonal variations of AVS–SEM in sediments and the degree to which the sampling locations in the cove properly represented its benthic conditions.

Exposures and effects on eelgrass in Clark Cove were evaluated from measurements on the bed on the northeastern edge

Table 6. Summary of risk to assessment endpoints in Clark Cove, Maine, USA

| Assessment endpoint | Evidence of effect[a] | Evidence of exposure[b] | Magnitude of risk | Confidence in conclusions |
|---|---|---|---|---|
| Pelagic | Potential/M | Low/M | Low | Medium |
| Epibenthic | No/M | Elevated/M | Low | Medium |
| Benthic | No/H | Elevated/M | Low | High[c] |
| Eelgrass | Potential[d]/M | Elevated/M | Intermediate | Medium |
| Salt Marsh | No/M | Elevated/M | Low | Medium |
| Avian[e] | | Negligible/M | Negligible | Medium |

[a] Entry = evidence of effect/endpoint weight (H = high, M = medium, L = low).
[b] Entry = evidence of exposure/endpoint weight (H = high, M = medium, L = low).
[c] High concordance between highly weighted measures.
[d] Eelgrass was absent within Clark Cove.
[e] Risk of dietary exposure for Portsmouth Harbor, New Hampshire.

of the cove. Chemicals in leaf and root tissue and in bulk sediment were higher than their respective backgrounds (Fig. 3D). No effects were found in any of the measures of morphology or density measures made in plants sampled from Clark Cove. Because inner Clark Cove contained no eelgrass beds, however, we used our professional judgment to interpret the absence of eelgrass to be a potential effect that outweighed the measurements of no effect (Fig. 3D). Since most of Clark Cove is too deep to support eelgrass anyway, the affected area is probably limited to suitable eelgrass habitat along the fringes of the inner cove. The reason for the absence of eelgrass in inner Clark Cove was unknown, and the spatial extent of suitable eelgrass habitat in Clark Cove was not measured. We used professional judgment to reach a medium weight of potential effect and of elevated exposure to eelgrass receptors (Fig. 3D).

The weight of evidence for exposure and effects to the salt marsh community in Clark Cove indicated medium weight of no effect and elevated exposure (Fig. 3E). While some of the measures suggested potential effects, most indicated none (Fig. 3E). Chemical concentrations in *Spartina* leaf tissues suggested negligible exposure, but concentrations in bulk sediment suggested high exposure (Fig. 3E). While the weight of evidence suggested no effect to the salt marsh community, we noted that the low, middle, and high zones of the marsh differed greatly. Part of this was probably natural heterogeneity such as high numbers of barnacles on rocks increasing the number of animal taxa in the low marsh. Even though the marsh had well-developed substrata of peat, that zone was small; the western two thirds of the seaward edge of the marsh had only a narrow band of tall *S. alterniflora* present. In addition, some patches of short *S. alterniflora* communities were not sampled [7,15].

Negligible exposure to avian receptors was concluded because the calculated hazard index was less than two for all dietary pathways (Fig. 3F). All hazard quotients were less than 1.0 for all species and food items except for tDDx (hazard quotient = 1.26) based on a diet of 100% winter flounder by herring gulls. It was assumed that most feeding scenarios will not reach the level of exposure predicted in the models and therefore potential risks would be lower that those modeled. In light of these conditions, we assumed that there was negligible risk of exposure to upper food-chain species [7].

### *Interpretation of risk*

The magnitude and confidence of risk in Clark Cove (Table 6) was defined by combining the evidence for exposure and effect. The evidence of effect and exposure were obtained from the centroids plotted for each assessment endpoint, which gave the level of exposure or effect and its associated endpoint weight (e.g., potential/*M*). The magnitude of risk was obtained from the combination of exposure and effects evidence defined in Table 2. The confidence in conclusion reflected the average endpoint weights obtained for evidence of effects and exposure, the degree of concurrence among the endpoint weights for evidence of effects and exposure, the degree of concurrence between conclusions regarding magnitudes of exposure and effect, and professional judgment used to qualify conclusions. For example, we had medium confidence of low risk to pelagic receptors in Clark Cove because, while there was medium weight of potential effect, there was medium weight of low exposure. Similarly, we had high confidence of low risk to benthic receptors in Clark Cove because there was high weight of no effect with medium weight of elevated exposure (Table 6).

### *Risks from environmental media*

By relating risk back to exposure to surface water and sediment, the risk in Clark Cove from its environmental media was estimated. We concluded that there was medium confidence of low risk from surface water and high confidence of low risk from sediment to ecological receptors. We also concluded that there was negligible risk of dietary exposure to avian receptors. We qualified these conclusions with the following caveats. The evidence of bioaccumulation in mussels is probably related to surface water exposure, and the elevated concentrations in tissues of juvenile lobsters are related to sediment exposure. We also recognized that resuspended fine-grained sediment in areas like Clark Cove might contribute to the risks from exposure to surface water.

## DISCUSSION

Multiple measures of exposure and effect were obtained from ecological studies of the estuary. Unfortunately, however, the various measurement data differed in uncertainty, in reliability for suggesting harm to the assessment endpoint, and in the degree of harm predicted for the endpoint. These differences made the results very difficult to interpret. Rather than relying on ad hoc judgment for interpreting risk, all the available data were systematically evaluated to determine whether a result added weight to the conclusion of risk or added weight to the conclusion of no risk.

Because no single measure can satisfactorily determine risk, multiple lines of evidence were used. This weight-of-evidence analysis allowed us to derive the risk estimate and confidence

levels upon which the final conclusions were based. This systematic way of reaching conclusions is intended to be transparent and produces an objective and consistent interpretation of the results. By formulating the conclusions within the context of the decision-making process, we used the results from the weight-of-evidence analysis to develop conclusions about risk that supported risk management decisions at the shipyard.

*Endpoint weights*

The procedure for weighting endpoints can be thought of as a means for ranking the relative uncertainty and reliability of the measures used in the risk assessment. We weighted measures high whose data were less uncertain and more reliable for assessing harm to the endpoints. We weighted measures medium and low whose data were more uncertain and less reliable.

We assumed that each assessment endpoint was equally important in the overall function of the ecosystem. Within each assessment endpoint, weights were assigned to the various measures that relate independently to it. (For example, data on mussels and lobsters provide information on the epibenthic assessment endpoint but mussels and lobsters may be affected differently by stressors.) The weighting scheme helped balance the importance of each factor with the quality and usefulness of the data. Assuming equal weights, finding no effect from one measure is just as important as finding a potential or probable effect from another measure.

We felt that quality of data was particularly important to the categories of measurement attributes. Data of low quality should not be included in the weight-of-evidence analysis because they could lead to spurious conclusions. There is a great deal of difference between low-quality data and low strength of association or study design. Low strength of association or low study design simply means that less weight will be assigned to the result for purposes of interpretation, whereas low-quality data cannot be interpreted because they are unreliable (e.g., analytical chemistry data that do not meet minimum quality control/quality assurance objectives). We also considered the possibility that poor data could eliminate important measures or that superior data could increase the effect of less important measures on conclusions, which could be a problem when including measures that were not related to the assessment endpoint being evaluated. We avoided this problem by not including the latter kind of measures. We were also careful to define the relationships between the measures and the assessment endpoints when weighting the endpoints.

The strength-of-association category of attributes (degree of association, response to stressor, and utility of measure) can be considered intrinsic properties of the measure, which means their weights will depend on how sensitive and robust the measure is in assessing harm to the assessment endpoint. To increase the weight of a measure in this category, one must demonstrate the relationship between the endpoint and the measure, establish sensitive benchmarks, and improve the scientific basis for inferring harm. We found that, for most measures, the study-design attributes had the greatest opportunity to improve the overall weight of the endpoint. Low weights were usually assigned to study designs that contained too few intervals of sampling (i.e., that lacked temporal representativeness) and whose measures could not differentiate stressor responses from natural stochasticity. Improvements in the design and execution of the studies that obtained the measurement data could yield more highly weighted measures for inferring risk.

Most measures evaluated (Table 5) were weighted high for data quality, medium for strength of relationship, and medium for study design, which indicated moderate to low uncertainties in the measurement data. This was because the studies that assessed risk were site specific, were directed at specific ecological components and receptors within the area, used standardized sampling and analysis, complied with appropriate procedures for quality control and quality assurance, and provided measurements that applied to the assessment endpoint.

One of the main limitations of the method is that the measurement endpoints must be representative of the assessment endpoints and that the results obtained from the measures must be indicative of ecological risk. Although the measurement endpoints were weighted after the risk-assessment studies were completed, the weighting exercise provided a way to reach a consensus and formulate conclusions. This allowed us to keep the characterization of risk focused on the data. Weighting the measurement endpoints during problem formulation [5] would result in the selection and design of studies that could result in more clearly described risks.

*Evidence of risk*

Data for measures of effect were evaluated to determine whether the outcome added to conclusions that effects on endpoints were or were not evident. Measures of exposure were evaluated relative to the conclusions that exposure would or would not cause an effect. In this sense, measures of exposure with benchmarks of effects (e.g., concentrations in surface water, sediment, or prey) could be evaluated regarding whether the benchmark was exceeded and if so by how much. When benchmarks were not available (e.g., fucoid algae residues), we had to compare the results from the areas of concern with data from reference areas. In turn, reference areas were used to evaluate effects relative to pristine areas and to other areas of the estuary. Reference data used for comparison carried the same relative weight as the measure being evaluated. The appropriateness of the reference data was evaluated as part of the study design contribution to the measure's endpoint weight.

While interpreting the weight of evidence, we found that we had to think in terms of the full body of evidence. This is especially important when dealing with equivocal results. For example, the weight of evidence for effects to the pelagic endpoint contained conflicting evidence (Fig. 3A). Although toxicity to sea urchin fertilization indicated a probable effect, phytoplankton biomass and growth of deployed mussels indicated no effect. Because these measures had similar weights, no clear, unequivocal conclusion could be identified. Therefore, potential effect was the only accurate description for the pelagic endpoint. Here the weight of evidence balanced the evidence rather than tipping the scale, as in a court of law where the greater amount of evidence can sustain a verdict [29].

Alternatively, one might propose that any evidence of an effect (or exposure) would fix the conclusion. We rejected this reasoning because additional measures will increase the confidence in the conclusion and decrease the chance of its being swayed by outliers or spurious results. Additionally, one may be certain of the results of one measure but less confident about conclusions drawn from one line of evidence. Individual measures are uncertain, but multiple lines of evidence reinforce confidence [30]. Basing a conclusion on many lines of evidence will increase confidence in the conclusion even though the un-

certainties of the individual measures increase the overall uncertainty.

Additional measures may also dilute the evidence for a real effect. Again, using the evidence of effect to pelagic receptors in Clark Cove (Fig. 3A), the probable effect indicated by toxicity to sea urchin fertilization could only be proposed by ignoring the no effect suggested by phytoplankton biomass and growth in deployed mussels. Since the assessment endpoint was the health and vitality of pelagic species, toxicity to sea urchin fertilization was only a partial indicator for the pelagic species that broadcast their sperm and larvae. The conflicting results affected our confidence. Clearly, if all the lines of evidence were in agreement, we would have much greater confidence in the conclusion rendered. Because we agreed that no single measure is conclusive for determining ecological risk [30], judging all the data bolsters the conclusions and results in more accurate assessments of risk [1,31–34].

We reserved the right to invoke our professional judgment, as we did when concluding a potential effect to eelgrass in Clark Cove (Fig. 3D), if the balance of evidence (centroid) suggested conclusions that were contrary to our overall understanding. For the most part, this was rarely necessary because we felt that the weight-of-evidence analysis accurately captured the situation at the sites. By taking care to weight the measures accurately and objectively, we developed a ranking system that contained the strengths and weaknesses of the measures. By systematically analyzing the weight of evidence and developing a consensus among ourselves, we strove to eliminate personal or professional bias as much as possible.

One of the objectives of every risk assessment is to clearly communicate the results and the major factors that influenced them. In characterizing risks for each area of concern, we carefully qualified the conclusions by describing their rationale, which consisted of the major sources of confidence and of uncertainty. This information is valuable to risk managers and other stakeholders because it makes the process of characterizing risk easier to understand, more explicit, and hopefully more widely acceptable. Even though some might not agree with our conclusions, the process clearly shows how we derived them.

### *Risks from environmental media*

Believing that contaminants released in the estuary would follow the hypothesized pathways of exposure (Fig. 2), we expected the assessment endpoints to respond differently to different pathways. While it is difficult to separate the exposures from water, sediment, and food, we assumed that the measures used to evaluate the pelagic and benthic assessment endpoints would be more affected by surface water and sediment, respectively. We also assumed that the measures used to assess the epibenthic, eelgrass, and salt marsh assessment endpoints were equally affected by exposure from water and sediment. The avian endpoint was a special case because we did not have any measures of effect to avian consumers for evaluating the risk. Since the measures for the avian endpoint were modeled from dietary exposure, we could draw conclusions only about the potential risk of dietary exposure to avian receptors. The risk from the exposure media (water and sediment) was based on the risks determined for each assessment endpoint weighted by the predominant route of exposure (Table 3, Fig. 2).

Even though the weighting involved numerical calculations, the analysis was really qualitative. The calculations only helped us to synthesize all the qualitative evaluations to that point. By linking all the evaluations (weighting the endpoints, evaluating the evidence of exposure and effects, determining the magnitude of risk, and attributing the risk to the various media) into one systematic procedure, it is possible to see how a particular judgment will affect the final conclusion. Accordingly, the manner in which we reached the conclusions is transparent. If new information becomes available, we can quickly determine how it would change our conclusions. By providing clear descriptions of risks and how they were derived, the conclusions were intended to support the decisions that are part of managing risk at a site. For low, intermediate, or high risk, development of preliminary remediation goals and feasibility study are recommended and, in cases of high risk, removal actions may be warranted. Risk-management decisions should also consider the degree of confidence in the conclusions. Low confidence suggests that additional information could change the conclusion; high confidence suggests the opposite. The background (ambient) risk should also be considered to ensure that remedies would not be nullified by larger scale problems. In this sense, the magnitude of risk may play an important role in setting the priorities for cleanup.

## CONCLUSION

Although the weight-of-evidence approach and the individual measures from which the risks were evaluated carried their own uncertainties, the conclusion becomes stronger as more information is used [1–2,30–34]. Separating exposure and effects measures and assigning weights to individual measures allowed us to tie together diverse data from multiple stressors and effects, keep track of the basis for the risk estimate, and incorporate uncertainty into the conclusions about risk. Improvements to the methodology recommended by Menzie et al. [5] included plotting the outcomes of exposure and effects measures and the centroid to visualize the weight of evidence, defining risk based on the exposure and effects evidences, relating the estimate of risk back to the exposure media, and explicitly expressing the confidence in conclusions. This case study showed the utility of the procedures recommended by Menzie et al. [5] and demonstrated that multiple lines of evidence can be assigned different weights to develop conclusions about risk in a manner that was clearly defined, objective, consistent, and did not rely solely on professional judgment. We believe that by following the weight-of-evidence analysis described here, the strengths and weaknesses of ecological risk assessments can be incorporated into the conclusions about risks and the decisions that will help manage them.

*Acknowledgement*—This work was supported by a cooperative research and monitoring agreement between the U.S. Navy and the U.S. EPA. Supporting work was conducted by the University of New Hampshire, the University of Rhode Island, Science Applications International Corporation, Battelle Marine Sciences Laboratory, and Ceimic Corporation. We appreciate the contributions from M. Cassidy, J. Conroy, F. Evans, and N. Beardsley and critical reviews provided by I. McLeod, D. Brown, M. Pilson, K. Rahn, B. Brown, the Portsmouth Naval Shipyard Restoration Advisory Board, and two anonymous reviewers. This work was funded by the U.S. Navy Marine Environmental Support Office, Northern Division Naval Facilities Engineering Command, and the U.S. EPA Office of Research and Development. This publication reflects the personal views of the authors and does not suggest or reflect the policy, practices, programs, or doctrine of any U.S. governmental agency. Mention of trade names or commercial products does not constitute either endorsement or recommendation for use.

## REFERENCES


1. Cook RB, Suter GW II, Sain ER. 1999. Ecological risk assessment in a large river–reservoir: 1. Introduction and background. *Environ Toxicol Chem* 18:581–588.
2. Culp JM, Lowell RB, Cash KJ. 2000. Integrating mesocosm experiments with field and laboratory studies to generate weight-of-evidence risk assessments for large rivers. *Environ Toxicol Chem* 19:1167–1173.
3. Anderson BS, et al. 2001. Sediment quality in Los Angeles Harbor, USA: A triad assessment. *Environ Toxicol Chem* 20:359–370.
4. McGee BL, Fisher DJ, Yonkos LT, Ziegler GP, Turley S. 1999. Assessment of sediment contamination, acute toxicity, and population viability of the estuarine amphipod *Leptocheirus plumulosus* in Baltimore Harbor, Maryland, USA. *Environ Toxicol Chem* 18:2151–2160.
5. Menzie C, et al. 1996. Special report of the Massachusetts weight-of-evidence workgroup: A weight-of-evidence approach for evaluating ecological risks. *Human Ecol Risk Assess* 2:277–304.
6. Johnston RK, Munns WR Jr, Mills LJ, Short FT, Walker HA, eds. 1994. Estuarine ecological risk assessment for Portsmouth Naval Shipyard, Kittery, Maine: Phase I: Problem formulation. Technical Report 1627. Naval Command, Control, and Ocean Surveillance Center, Research Development Test and Evaluation Division, San Diego, CA, USA.
7. Marine Environmental Support Office. 2000. Estuarine ecological risk assessment for Portsmouth Naval Shipyard, Kittery, Maine. Final Report, Volumes I and II. Marine Environmental Support Office—East Detachment, Space and Naval Warfare Systems Center, Narragansett, RI, USA.
8. U.S. Environmental Protection Agency. 1991. Resource Conservation and Recovery Act facility investigation proposal with conditions. Letter from U.S. EPA Region I to Commanding Officer, Portsmouth Naval Shipyard, January 15, 1991. Boston, MA, USA.
9. U.S. Environmental Protection Agency. 1994. National priorities list for uncontrolled hazardous waste sites: Final rule. *Fed Reg* 59:27989–28024.
10. New Hampshire Office of State Planning. 1997. 1996 population estimates of New Hampshire cities and towns. Concord, NH, USA.
11. Chadwick J. 1993. Application of the water quality model WASP3 for the Great Bay Estuary. MS thesis. University of New Hampshire, Durham, NH, USA.
12. Pavlos JR. 1994. Application of a one dimensional pollutant fate model for the Great Bay Estuary. MS thesis, University of New Hampshire, Durham, NH, USA.
13. Ward LG. 1995. Sedimentology of the Lower Great Bay/Piscataqua River Estuary. Contribution 314. Final Report. Jackson Estuarine Laboratory, University of New Hampshire, Durham, NH, USA.
14. Short FT, ed. 1992. *The Ecology of the Great Bay Estuary, New Hampshire and Maine: An Estuarine Profile and Bibliography.* NOAA Coastal Ocean Program, The Great Bay National Estuarine Research Reserve, Durham, NH, USA.
15. Burdick DM. 1994. Population characteristics of the salt marsh communities on and around the Portsmouth Naval Shipyard. Contribution 300. Final Report. Jackson Estuarine Laboratory, University of New Hampshire, Durham, NH, USA.
16. Naval Energy and Environmental Support Activity. 1983. Initial assessment study Naval Shipyard Portsmouth. NEESA 13-032. Port Hueneme, CA, USA.
17. McLaren/Hart Environmental Engineering. 1992. RCRA facility investigation for Portsmouth Naval Shipyard. Draft Final Report. Albany, NY, USA.
18. Halliburton NUS. 1995. On-shore feasibility study report for Portsmouth Naval Shipyard. Draft Report. Wayne, PA, USA.
19. Brown and Root Environmental. 1997. Draft site management plan for the Portsmouth Naval Shipyard. Pittsburgh, PA, USA.
20. U.S. Environmental Protection Agency. 1992. Framework for ecological risk assessment. 630/R-92/001. Final Report. Washington, DC.
21. U.S. Environmental Protection Agency. 1998. Guidelines for ecological risk assessment. 630/R-95/002f. Final Report. Washington, DC.
22. Suter GW II. 1993. *Ecological Risk Assessment.* Lewis, Boca Raton, FL, USA.
23. Long ER, MacDonald DD, Smith SL, Calder FD. 1995. Incidence of adverse biological effects within the ranges of chemical concentrations in marine and estuarine sediments. *Environ Manag* 19:81–97.
24. Strobel CJ, Buffum HW, Benyi SJ, Petrocelli EA, Reifsteck DR, Keith DJ. 1993. Virginian Province demonstration report: EMAP-Estuaries: 1990. 620/R-93/006. Environmental Monitoring and Assessment Program. Narragansett, RI, USA.
25. Windom HL, et al. 1989. Natural trace metal concentrations in estuarine and coastal marine sediments of the SE United States. *Environ Sci Technol* 23:314–320.
26. Di Toro DM, Mahony JD, Hansen DJ, Scott KJ, Hicks MB, Mayr SM, Redmond MS. 1990. The toxicity of cadmium in sediments: The role of acid volatile sulfide. *Environ Toxicol Chem* 9:1487–1502.
27. Di Toro DM, et al. 1991. Technical basis for establishing sediment quality criteria for nonionic organic chemicals using equilibrium partitioning. *Environ Toxicol Chem* 10:1541–1583.
28. Sample B, Opresko D, Suter GW II. 1996, Toxicological benchmarks for wildlife: 1996 revision. ER/TM-86/R3. Oak Ridge National Laboratory, Oak Ridge, TN, USA.
29. Black HC, Nolan JR, Nolan-Haley JM. 1990. *Black's Law Dictionary.* West Publishing, St. Paul, MN, USA.
30. Suter GW II. 1999. Lessons for small sites from assessments of large sites. *Environ Toxicol Chem* 18:579–580.
31. Suter GW II, Barnthouse LW, Efroymson RA, Jager H. 1999: Ecological risk assessment in a large river–reservoir: 2. Fish community. *Environ Toxicol Chem* 18:589–598.
32. Jones DS, Barnthouse LW, Suter GW II, Efroymson RA, Field JM, Beauchamp JJ. 1999. Ecological risk assessment in a large river–reservoir: 3. Benthic invertebrates. *Environ Toxicol Chem* 18:599–609.
33. Sample BE, Suter GW II. 1999. Ecological risk assessment in a large river–reservoir: 4. Piscivorous wildlife. *Environ Toxicol Chem* 18:610–620.
34. Baron LA, Sample BE, Suter GW II. 1999. Ecological risk assessment in a large river–reservoir: 5. Aerial insectivorous wildlife. *Environ Toxicol Chem* 18:621–627.

## APPENDIX

Measurement attributes, evaluation criteria, and weighting score values used to weight measures of exposure and effects

| Attribute | Evaluation criteria | Weighting score[a] |
|---|---|---|
| Data quality | Did data from the measure attain data quality objectives for sensitivity, precision, accuracy, completeness, representativeness, and comparability? | H = data met all data quality objectives<br>M = one data quality objective not met<br>L = data failed to meet two or more data quality objectives; not included in the risk characterization |
| Strength of association | Is there a biological linkage between the measure and the assessment endpoint, a correlation between the measure's response and stressor levels, and is there a scientific basis for using the measure to judge environmental harm? | H = the measure is equivalent or similar to the assessment endpoint, a statistically significant correlation exists between stressor levels and the measure's response, there is a high to moderate scientific basis for inferring environmental harm, and sensitive benchmarks are available<br>M = the measure is linked to the assessment endpoint but the level of biological organization is different, there is a quantitative relationship between measurement response and stressor levels, and although benchmarks may not be available, there is a moderate scientific basis for inferring harm<br>L = the measure is affected by factors unrelated to stressor levels, a correlation between stressor levels and measurement response is expected but not demonstrated, benchmarks are not available, and a relationship between the measure has been suggested or is expected but the scientific basis for inferring harm is weak or lacking |
| Study Design | Was the study designed to account for (1) specifics of the site, (2) spatial variation, and (3) temporal changes; was the measure (4) sensitive to changes due to stressor levels; was the measure able to (5) provide quantitative data, and was the measure (6) reproducible, applicable, suitable, and acceptable for assessing environmental harm? | H = the data obtained from the measure met five or six of the evaluation criteria<br>M = the data obtained from the measure met four or five of the six evaluation criteria<br>L = the data obtained from the measure was unable to meet three or more of the six evaluation criteria |

[a] Measures were assigned an endpoint weight score of high (H), medium (M), or low (L) relative to their ability to assess harm to the assessment endpoint.